Zautomatyzowana analiza afektywna i multimodalne modele rozpoznawania emocji

Modele teoretyczne, percepcja twarzy i mowy, sygnały fizjologiczne, architektury Edge AI,

Streszczenie

Przedmiotem raportu jest analiza systemów estymacji zachowań ekspresyjnych i stanów afektywnych z obrazu, dźwięku, tekstu oraz sygnałów fizjologicznych. Omówiono podejścia kategoryczne, wymiarowe i opisowe, kodowanie Action Units według FACS, przetwarzanie sekwencyjne i fuzję multimodalną. Przedstawiono charakterystyki 23 odrębnie nazwanych rozwiązań, zachowując informacje z 26 wpisów zestawienia, w tym powtórzenia niektórych architektur i wariantów ich opisu. Zestawiono wyniki na zbiorach RAF-DB, AffectNet, FER2013, DEAP, AMIGOS, MELD, CMU-MOSEI, RAVDESS, WESAD i innych, nie utożsamiając ich miar. Zaproponowano architekturę Edge AI, kryteria odtwarzalnej walidacji, zasady bezpieczeństwa danych oraz interpretacji rezultatów. Rezultaty z badań wtórnych oznaczono jako raportowane; ich bezpośrednia porównywalność zależy od identycznego podziału danych, definicji zadania oraz metryki. Szczególną uwagę poświęcono ograniczeniom inferencji o subiektywnych emocjach z obserwowalnych sygnałów.

Słowa kluczowe: affective computing; automatyczna analiza twarzy; FACS; EEG; PPG; EDA; rozpoznawanie emocji; multimodal fusion; Edge AI; walencja; pobudzenie; walidacja międzyosobnicza.

1. Zakres badań i ograniczenia interpretacji

Informatyka afektywna bada metody pozyskiwania, opisu i modelowania sygnałów współwystępujących z reakcjami emocjonalnymi. Na potrzeby modelowania komputerowego konieczne jest rozdzielenie trzech obiektów: obserwowalnej ekspresji (np. jednostek działania mięśni twarzy), etykiety przypisanej przez obserwatora oraz subiektywnego stanu deklarowanego przez badanego. Utożsamienie tych poziomów prowadzi do nadinterpretacji danych [1–4].

System, który rozpoznaje uśmiech lub określoną konfigurację AU, nie identyfikuje automatycznie emocji radości. Mimika, gest, ton wypowiedzi i parametry autonomiczne zależą od sytuacji, kultury, stanu zdrowia, wysiłku, różnic indywidualnych i warunków pomiaru. Ocena naukowa powinna odnosić się do operacyjnej definicji zmiennej docelowej i sposobu pozyskania etykiet. Wskaźniki fizjologiczne również nie stanowią jednoznacznego odczytu emocji; mierzą zmiany aktywności układów biologicznych, które mogą mieć wiele przyczyn [4].

2. Modele teoretyczne emocji i konstrukcja etykiet

Podejście kategorialne. Klasy emocji reprezentuje się jako rozłączne etykiety, np. radość, smutek, strach, złość, wstręt lub zaskoczenie. Jest to wygodne dla modeli klasyfikacyjnych, ale wymaga zdefiniowania neutralności, współwystępowania stanów i sposobu etykietowania niejednoznacznych obserwacji.

Podejście opisowe i FACS. Facial Action Coding System rozdziela obserwowalne jednostki działania mięśni twarzy (Action Units) od interpretacji emocjonalnej. AU można modelować jako zmienne binarne lub natężenia porządkowe. Jest to bardziej falsyfikowalny i transparentny cel niż automatyczne wnioskowanie o ukrytym stanie emocjonalnym [2].

Podejście wymiarowe. W modelu walencja–pobudzenie afekt opisuje położenie na osiach ciągłych. Pozwala to analizować nasilenie i dynamikę, ale wymaga określonej skali, sposobu normalizacji, strategii anotacji i miar regresyjnych takich jak RMSE lub concordance correlation coefficient (CCC) [3].

Podejście deskryptywne. Modele multimodalne mogą generować językowy opis zaobserwowanych sygnałów oraz interpretację kontekstową. Generowany tekst nie jest bezpośrednim pomiarem emocji i musi być walidowany względem anotacji, jakości dowodów, stabilności i ryzyka konfabulacji.

Rama	Obiekt pomiaru / właściwe miary
Kategoryczna	Klasa etykiety; macro-F1, balanced accuracy, confusion matrix.
FACS / AU	Obecność i intensywność ruchu mięśni; AU-F1, ICC, zgodność anotatorów.
Wymiarowa	Walencja/pobudzenie; RMSE, MAE, CCC i kalibracja predykcji.

Opisowa	Opis kontekstowy; oceny ekspertów, zgodność dowodowa, bezpieczeństwo interpretacji.

3. Czasowa charakterystyka odpowiedzi afektywnych

Ocena stanów afektywnych wymaga jawnego określenia horyzontu czasowego. Modele operujące na pojedynczych klatkach nie opisują trajektorii wyrazu twarzy ani zmian prozodii. Z kolei bardzo długie okna mieszają reakcje na różne bodźce i komplikują ustalenie etykiet. Poniższe przedziały są przykładową organizacją analizy, nie uniwersalną klasyfikacją biologiczną.

Skala	Okno	Obserwowane zmienne
Krótkoterminowa	0,5–2 s	Ruchy mięśni twarzy, krótkie zdarzenia akustyczne, reakcje na bodźce.
Średnioterminowa	2–10 s	Trajektorie ekspresji, zmiany prozodii, wypowiedzi, odpowiedzi autonomiczne.
Dłuższa	10–60 s i więcej	Trendy EDA/HRV i kontekst interakcji; nie utożsamiać z pomiarem nastroju.

4. Tor wizyjny: zautomatyzowana analiza twarzy

Typowy potok rozpoczyna detekcja twarzy i ustalenie jej położenia w czasie. Dalej wykonuje się normalizację geometrii (wycinek, skalowanie, korektę rotacji) i ekstrakcję cech. Reprezentacje geometryczne kodują landmarki oraz ich relacje, reprezentacje wyglądowe obejmują teksturę i lokalne wzorce obrazu, a reprezentacje ruchu wykorzystują sekwencje klatek lub przepływ optyczny. Klasyczne podejścia, takie jak AAM, CLM, HOG czy SIFT, zachowują znaczenie jako punkty odniesienia i narzędzia diagnostyczne, podczas gdy współczesne modele CNN i transformery integrują ekstrakcję cech z klasyfikacją.

Wynik klasyfikacji zależy od jakości lokalizacji twarzy, zasłonięcia jej części, obrotu głowy, oświetlenia, rozdzielczości, artefaktów kompresji i długości sekwencji. Podczas oceny należy raportować odsetek klatek odrzuconych przez detektor, ponieważ podawanie jakości wyłącznie dla poprawnie wykrytych twarzy może istotnie zawyżać skuteczność całego systemu.

5. Tor mowy i tekstu

Sygnał akustyczny. Cechy obejmują wysokość tonu, energię, dynamikę głośności, tempo, pauzy, formanty i reprezentacje wyuczone. Analiza emocji w mowie wymaga kontroli treści, języka, typu mikrofonu oraz tożsamości mówcy. Mikrofoniczny szum i pogłos mogą powodować większą zmianę cech niż sam bodziec eksperymentalny.

Treść językowa. Modele analizują znaczenie, relacje semantyczne, negację oraz kontekst dialogu. Sarkazm i wypowiedzi sprzeczne z prozodią motywują fuzję tekstowo-akustyczną. Klasyfikacja tekstu i analiza głosu powinny być oceniane osobno przed ich integracją.

Systemy konwersacyjne. Rozwiązania typu EVI 3 i Raven służą innym zadaniom niż klasyfikatory na benchmarkach. Brak publicznej, porównywalnej metryki rozpoznania emocji nie pozwala traktować ich jako najwyżej ocenionych modeli; podawane czasy interakcji mają charakter deklaracji produktu i wymagają testów całego toru.

6. Sygnały fizjologiczne oraz jakość pomiaru

EEG. Może dostarczać cech widmowych, przestrzennych i czasowych powiązanych z reakcją na bodźce. Interpretacja wymaga rejestracji artefaktów ruchowych i ocznych, kontroli impedancji, referencji, częstotliwości próbkowania i separacji osoby badanej pomiędzy zbiorami treningowymi i testowymi.

EDA/GSR. Wskaźnik aktywności elektrodermalnej związany z układem współczulnym; typowo analizuje się składową toniczną i reakcje fazowe. Nie jest miarą konkretnej emocji i zależy między innymi od termiki i ruchu.

ECG i PPG. Dają informacje o czynności serca i, pod warunkiem odpowiedniej jakości, jej zmienności. Obliczenie HRV wymaga wykrycia interwałów i kontroli błędów; pomiary PPG są podatne na artefakty ruchowe i nie są równoważne diagnostycznej elektrokardiografii.

Dodatkowe dane. EMG, temperatura skóry, IMU i parametry otoczenia umożliwiają rozróżnianie artefaktów od zmian interesujących badacza oraz lepszą kontrolę kontekstu. Konieczne jest synchronizowanie zegarów i dokumentowanie niepewności opóźnień.

7. Architektury przetwarzania wielomodalnego

Fuzja wczesna konkatenatuje lub łączy znormalizowane reprezentacje wejściowe przed klasyfikatorem. Fuzja pośrednia wykorzystuje enkodery specyficzne dla modalności, a następnie attention lub cross-modal transformer. Fuzja późna agreguje predykcje już po klasyfikacji, co upraszcza obsługę brakujących modalności, ale może utracić zależności nieliniowe między sygnałami. Dla badań niezbędny jest eksperyment ablacyjny obejmujący każdy tor osobno oraz fuzję.

Hybryda CNN–BiLSTM–Transformer łączy ekstrakcję lokalnych cech, modelowanie zależności w sekwencji i mechanizm uwagi. Dwukierunkowa LSTM może wykorzystywać przyszłe próbki, dlatego deklarowana praca w czasie rzeczywistym wymaga rozróżnienia trybu przyczynowego od wsadowej analizy offline. Cross-modal attention pozwala odzwierciedlać zgodność między mową, twarzą i fizjologią, lecz może również uczyć się artefaktów sprzętowych i korelacji pozornych.

Strategia	Zalety	Ograniczenia
Wczesna	Prosta implementacja, pełne wektory cech	Wymaga synchronizacji; wysokie wymiary; braki danych.
Pośrednia	Elastyczne kodery i cross-attention	Wysoki koszt treningu i ryzyko niejawnych skrótów.
Późna	Odporność na brak jednej modalności	Utrata interakcji na poziomie cech; zależy od kalibracji.
Adaptacyjna	Wagi jakości sygnałów zmieniają się w czasie	Konieczna walidacja niezawodności i mechanizmów abstencji.

8. Zbiory danych i nieporównywalność benchmarków

Zbiór	Dominujący charakter zadania	Zalecany raport wyników
RAF-DB / FER2013	Klasyfikacja ekspresji twarzy	Accuracy, macro-F1, podział i preprocess.
AffectNet-7/8	Ekspresje twarzy oraz warianty VA	Dokładna liczba klas, definicja neutralności, metryka.
DEAP / AMIGOS	Afekt i sygnały fizjologiczne	Schemat dychotomizacji VA, LOSO, ACC/CCC.
WESAD	Stres/warunki afektywne, wearables	Rozdział uczestników, etykiety i okna czasowe.
MELD / CMU-MOSEI	Emocje i multimodalny dialog	Macro-/weighted-F1, dane o sesjach i osobach.
RAVDESS / SEED-IV	Aktorzy mowy / sygnały neurofizjologiczne	Rozdział mówców/osób, setting i klasy.
Aff-Wild2 / MER-FG	Regresja VA / analiza multimodalna	CCC, RMSE albo właściwa metryka zadania.

Wynik accuracy = 94,2% na jednym zbiorze nie oznacza, że model jest skuteczniejszy od systemu uzyskującego 67% na innym. Różnice w liczbie klas, rozkładzie etykiet, sposobie prezentowania bodźców, podziałach osób i modelu anotacji mają wpływ na wynik. Wartości percentowe, macro-F1, CCC i RMSE są odmiennymi miarami, których nie wolno scalać w pojedynczy ranking.

9. Szczegółowy przegląd modeli i deklarowanych parametrów

W poniższych profilach zachowano nazwy modeli, opis wejść, architektur, deklarowanych rezultatów i ograniczeń sprzętowych. Pozycje powtórzone w opisach zestawiono łącznie, ale odnotowano ich warianty. Dane o modelach bez pełnego protokołu ewaluacji mają charakter informacji raportowanej i wymagają sprawdzenia w pierwotnej publikacji, repozytorium lub karcie modelu przed wykorzystaniem jako wynik eksperymentalny.

9.1. EmotionTFN / Emotion-Multi-Scale Temporal Fusion Network

Modalności	EEG, PPG, GSR, obraz, dźwięk
Architektura	Hierarchiczna uwaga czasowa, 1D CNN, dilated convolutions i self-attention

Wynik raportowany	94,2% na MELD (wartość raportowana); w alternatywnym opisie brak wartości
Wdrożenie / zasoby	Przeznaczenie IoT i systemy czasu rzeczywistego; wariant opisany jako INT8, <200 ms na Jetson Xavier NX, lecz bez pełnego protokołu.
Granice interpretacji	Zestawienie łączy opis modelu wielomodalnego z wersją ogólną; należy ustalić tożsamość architektury i konfiguracji.

9.2. Attention-Based CNN–BiLSTM–Transformer

Modalności	EEG, ECG/EKG, GSR/EDA, EMG oraz wariant PPG
Architektura	CNN dla cech lokalnych, BiLSTM dla sekwencji, Transformer i attention-based fusion
Wynik raportowany	88,2% DEAP; 89,5% AMIGOS (wartości raportowane)
Wdrożenie / zasoby	FP32 / mixed precision, PyTorch lub TensorFlow; istotny koszt pamięci i czasowej uwagi.
Granice interpretacji	Wariant figuruje także jako CNN–BiLSTM–Transformer; to jedna rodzina, nie dwa niezależne eksperymenty.

9.3. MemoCMT

Modalności	obraz, dźwięk, tekst
Architektura	Cross-Modal Transformer
Wynik raportowany	82,3% na CMU-MOSEI (raportowane)
Wdrożenie / zasoby	Model fuzji multimodalnej na poziomie cech; rozmiaru nie określono.
Granice interpretacji	Weryfikować, którą podmetodę i miarę oceniono.

9.4. JOYFUL

Modalności	dźwięk, obraz, tekst
Architektura	Graph Contrastive Learning + Transformer
Wynik raportowany	81,2% macro-F1 na MELD (raportowane)
Wdrożenie / zasoby	Fuzja pośrednia; rozmiar określony jakościowo jako średni.
Granice interpretacji	Macro-F1 nie jest porównywalne z accuracy innych modeli.

9.5. Wear-BioNet

Modalności	tętno, EDA, akcelerometr
Architektura	Ensemble CNN–GRU, późna fuzja
Wynik raportowany	84,5% WESAD (raportowane)
Wdrożenie / zasoby	Lekka architektura dla wearables i edge.
Granice interpretacji	Przed interpretacją konieczna jest kontrola splitu zależnego od uczestnika.

9.6. TMNet

Modalności	EEG, mowa
Architektura	Cross-modal Transformer

Wynik raportowany	88,7% SEED-IV (raportowane)
Wdrożenie / zasoby	Modelowanie korelacji elektroencefalografii i mowy.
Granice interpretacji	Sprawdzić, czy zestaw zawiera oba kanały w jednym protokole badania.

9.7. Edge-MER

Modalności	twarz, dźwięk
Architektura	Lekki CNN–LSTM
Wynik raportowany	79,4% RAVDESS (raportowane)
Wdrożenie / zasoby	Cel: zastosowania mobilne o małym opóźnieniu i poborze mocy.
Granice interpretacji	Bez pomiaru energii oraz opóźnienia end-to-end wniosek o oszczędności pozostaje hipotezą.

9.8. DAF-MHSCA

Modalności	obraz twarzy
Architektura	ResNet18, AFC, DAF i Multi-Head Spatial-Channel Attention
Wynik raportowany	92,49% RAF-DB; 67,72% AffectNet-7; 74,33% FER2013
Wdrożenie / zasoby	Zgłaszany rozmiar około 16,5 MB; FP16/FP32.
Granice interpretacji	Różne zbiory mają niejednolite trudności i zbiory klas.

9.9. FedMultiEmo / Federated CNN

Modalności	obraz twarzy, tętno, EDA, temperatura skóry
Architektura	CNN dla obrazu, Random Forest dla fizjologii; fuzja decyzji
Wynik raportowany	87,0% fuzja; 77,0% obraz; 74,0% fizjologia (raportowane)
Wdrożenie / zasoby	Opisany klient <200 MB RAM, INT8/FP16; możliwość federacyjnego uczenia.
Granice interpretacji	Wartości nie zawierają pełnego schematu FL, kosztu komunikacji ani ustawień prywatności.

9.10. Raven (Tavus API)

Modalności	wideo, mowa, język ciała
Architektura	Model percepcji kontekstowej w rozmowie wideo
Wynik raportowany	Brak liczby w porównywalnym benchmarku
Wdrożenie / zasoby	Architektura produktowa / cloud API z możliwym komponentem brzegowym.
Granice interpretacji	Nie klasyfikować jako SOTA bez wspólnej metryki.

9.11. MFER

Modalności	obraz twarzy
Architektura	Multiscale Facial Expression Recognition, global/local attention
Wynik raportowany	92,08% RAF-DB; 67,06% AffectNet-7; 73,45% FER2013

Wdrożenie / zasoby	FP32; koszt obliczeniowy określono jako średni.
Granice interpretacji	Różnice względem DAF-MHSCA są niewielkie na RAF-DB i wymagają przedziałów ufności.

9.12. DDAMFN

Modalności	obraz twarzy
Architektura	Dual-Directional Attention Mixed Feature Network
Wynik raportowany	91,35% RAF-DB; 67,03% AffectNet-7
Wdrożenie / zasoby	Brak podanego rozmiaru i precyzji.
Granice interpretacji	Nie wnioskować o przewadze wdrożeniowej bez pomiaru pamięci i FPS.

9.13. AffectGPT

Modalności	wideo, audio, tekst
Architektura	Multimodal Large Language Model z mechanizmem pre-fusion
Wynik raportowany	57,36% w wariancie S1 na MER-FG; inna pozycja bez metryki
Wdrożenie / zasoby	Wersja opisana jako ~7B parametrów, BF16/INT4; wysokie wymagania VRAM.
Granice interpretacji	Generowanie wyjaśnień należy oceniać także pod kątem oparcia w sygnale i błędów interpretacji.

9.14. AlexNet Variant

Modalności	obraz twarzy / AffectNet
Architektura	Lekka odmiana CNN AlexNet
Wynik raportowany	55,09% AffectNet-8
Wdrożenie / zasoby	Wariant mały, docelowo INT8/FP16, scenariusze mobile/web.
Granice interpretacji	AffectNet-8 i AffectNet-7 nie mogą być traktowane jako jeden benchmark.

9.15. Combined CNN–RNN

Modalności	sekwencje wideo
Architektura	Ekstraktor CNN + trójwarstwowa GRU
Wynik raportowany	RMSE 0,45; CCC 0,39 dla arousal na Aff-Wild2
Wdrożenie / zasoby	FP32; mała liczba parametrów względem dużych modeli sekwencyjnych.
Granice interpretacji	To regresja ciągła, nie klasyfikacja. Nie przeliczać RMSE/CCC na accuracy.

9.16. AffectNet+

Modalności	obraz wyrazu twarzy
Architektura	Głębokie sieci neuronowe
Wynik raportowany	90,4% bez jednoznacznej specyfikacji protokołu
Wdrożenie / zasoby	FP32/FP16.

Granice interpretacji	Wymaga identyfikacji eksperymentu, liczby klas oraz definicji zbioru.

9.17. EVI 3 (Hume AI)

Modalności	mowa, tekst
Architektura	Interfejs speech-to-speech / model językowo-mówiony
Wynik raportowany	Brak porównywalnej dokładności; deklaracja latencji <300 ms
Wdrożenie / zasoby	Integracje z modelami językowymi; platforma produktowa.
Granice interpretacji	Nie interpretować czasu odpowiedzi jako jakości rozpoznawania stanów afektywnych.

9.18. EMOE

Modalności	obraz, akustyka, język
Architektura	Mixture of Modality Experts; unimodal distillation
Wynik raportowany	Brak wartości benchmarku
Wdrożenie / zasoby	Wymaga akceleracji GPU/NPU dla cięższej fuzji.
Granice interpretacji	Parametry do ustalenia: liczba ekspertów, routing, koszt pamięci, odsetek brakujących modalności.

9.19. ArcFace (wariant na AffectNet)

Modalności	obraz twarzy
Architektura	Głowica z funkcją/strategią ArcFace
Wynik raportowany	87,6% (bez pełnego protokołu)
Wdrożenie / zasoby	Wysokie wymagania obliczeniowe według zestawienia.
Granice interpretacji	ArcFace jest metodą reprezentacji/uczenia cech; wyjaśnić dokładną implementację zadania emocji.

9.20. Poria i in., model 2015

Modalności	tekst, dźwięk, wideo
Architektura	Wielomodalny model uczenia głębokiego
Wynik raportowany	87,95% (bez pełnej identyfikacji zadania)
Wdrożenie / zasoby	Znaczne wymagania obliczeniowe i danych w opisie.
Granice interpretacji	Wynik historyczny, nieprzenoszalny na nowe zbiory.

9.21. Scaled-down EfficientNet

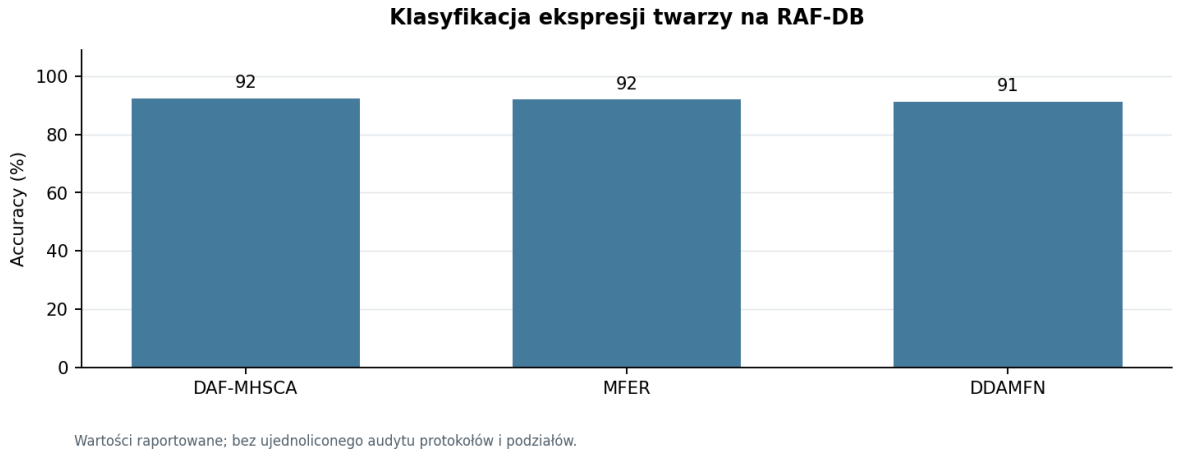
Modalności	obraz twarzy
Architektura	EfficientNet przeskalowany dla niskich zasobów
Wynik raportowany	61,3% (zbiór niejednoznaczny)
Wdrożenie / zasoby	Wariant mobilny / edge.
Granice interpretacji	Potrzebna pełna specyfikacja współczynnika skali i akceleratora.

9.22. ResNet50 (Video CNN)

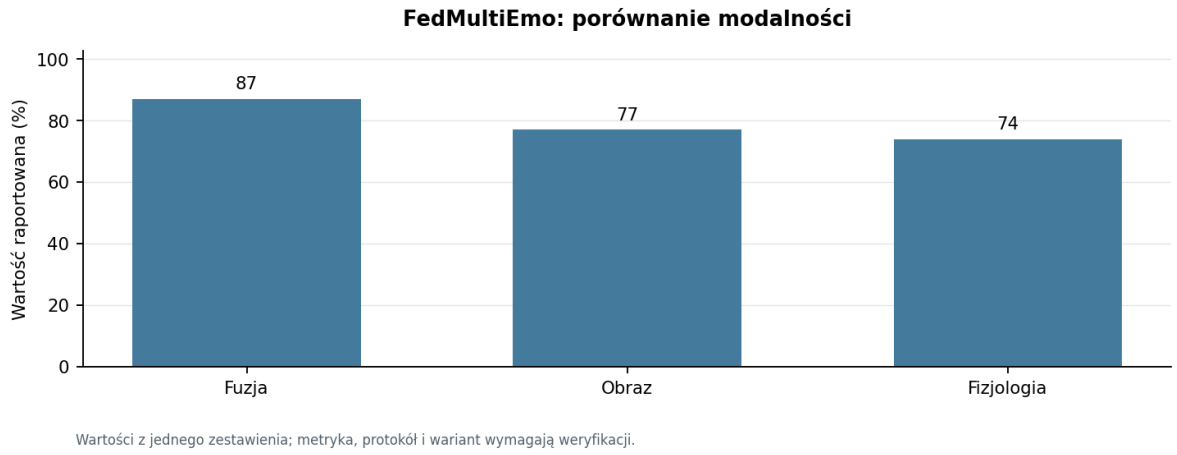
Modalności	wideo
Architektura	ResNet50 jako model wideo
Wynik raportowany	76% (bez pełnego protokołu)
Wdrożenie / zasoby	Deklaracja ~2,5 mln parametrów wymaga weryfikacji konfiguracji.
Granice interpretacji	Standardowy ResNet50 ma inną typową skalę; wymaga ustalenia zakresu liczonych parametrów.

9.23. C3D

Modalności	wideo
Architektura	Trójwymiarowe sploty czasowo-przestrzenne
Wynik raportowany	76% (bez pełnego protokołu)
Wdrożenie / zasoby	Deklaracja 0,33 mln parametrów dotyczy niejasnego wariantu.
Granice interpretacji	Bez szczegółów architektury nie wolno przenosić charakterystyki na standardową C3D.



Rysunek 1. Rezultaty klasyfikacji na RAF-DB: porównanie wyłącznie opisowe, ponieważ nie zweryfikowano identyczności wszystkich protokołów.



Rysunek 2. Wartości raportowane dla trzech torów FedMultiEmo; nie stanowią oszacowania efektu na innych zbiorach.

10. Wymagania urządzeń brzegowych i optymalizacja modeli

System Edge AI powinien zapewniać akwizycję sygnałów, synchroniczną segmentację, inferencję oraz telemetrię bez wymuszania transmisji surowego obrazu lub EEG do usług zewnętrznych. Dla platform klasy Jetson Orin oraz Rockchip

RK3588 istotne są dostępność NPU/GPU, zgodność runtime (TensorRT, RKNN, ONNX Runtime, OpenVINO), rzeczywista zajętość pamięci i stabilność termiczna. Czas pojedynczej inferencji to tylko fragment opóźnienia całkowitego.

Kwantyzacja INT8/FP16 może obniżać koszt pamięci i przyspieszać inferencję, ale jej wpływ na CCC, macro-F1 i podgrupy należy mierzyć po eksporcie modelu na konkretny akcelerator. Pruning, distillation i ograniczenie częstotliwości przetwarzania powinny być traktowane jako osobne interwencje eksperymentalne. Zmiana liczby ramek na sekundę może zmodyfikować reprezentację dynamiki mimiki, a skrócenie okna EEG zmienia estymowane cechy widmowe.

Warstwa	Zakres implementacji	Metryka kontrolna
Sensor / synchronizacja	RGB, audio, EEG/EDA/PPG, clock mapping	Jitter; utrata ramek; odchyłka czasu
Ekstrakcja	Detekcja twarzy, filtracja sygnałów, QC	Odsetek poprawnych próbek
Enkodery	CNN, CNN-RNN, transformer, feature model	P95 inferencji; RAM; waty
Fuzja	Wczesna, pośrednia, późna, adaptacyjna	Ablacja i odporność na braki
Decyzja	Etykieta / VA / abstencja	Macro-F1, CCC, kalibracja, OOD
Rejestracja	Wyłącznie niezbędne metadane, wersje modeli	Odtwarzalność i zgodność z prawem

Przykładowe zakresy 1–10 ms (edge), 10–100 ms (fog) i 100 ms–2 s (cloud) należy rozumieć wyłącznie jako ilustracyjne klasy opóźnień, a nie gwarantowane SLA. Konfiguracja sprzętowa, długość okna i liczba jednoczesnych modalności decydują o realnych wartościach; dla systemów multimodalnych często to akwizycja fizjologii, a nie inferencja sieci, ogranicza częstość aktualizacji.

11. Protokół walidacji porównawczej

Pytanie badawcze. Czy fuzja multimodalna poprawia wyniki w porównaniu z najsilniejszą modalnością osobno przy identycznym rozdziale uczestników i ograniczonym budżecie zasobów?

Jednostka niezależna. Uczestnik lub sesja w zależności od pytania; okien pochodzących od jednej osoby nie wolno dzielić losowo między trening i test, gdy skuteczność ma być uogólniana międzyosobniczo.

Warianty eksperymentalne. Osobne modele RGB, audio, tekst, EEG/EDA/PPG, ich pary oraz pełna fuzja. Dodatkowe eksperymenty: brak modalności, kompresja, niższa częstotliwość, hałas oraz przesunięcie domeny.

Metryki. Dla klas: macro-F1, balanced accuracy, AUROC tam, gdzie odpowiednia, confusion matrix i bootstrap CI po uczestnikach. Dla VA: CCC, MAE/RMSE, porównania Bland–Altman i jakość kalibracji. Dla wdrożenia: P50/P95/P99 opóźnienia, pamięć, energia, liczbę odrzuconych okien.

Niepewność. Różnice modeli porównywać na tych samych jednostkach testowych. Preferować bootstrap grupowy po uczestnikach i odpowiednie testy sparowane; raportować wielkość efektu i przedział ufności, nie tylko istotność statystyczną.

Ocenianie deskryptywne. Weryfikować, czy opisy wygenerowane przez MLLM odnoszą się do sygnałów faktycznie zaobserwowanych, czy dodają niepoparte interpretacje i czy model poprawnie wybiera abstencję przy niskiej jakości wejścia.

12. Bezpieczeństwo, prywatność i ograniczenia prawne

Dane z kamery, mowy i sygnałów fizjologicznych mogą umożliwiać identyfikację osoby lub ujawniać informacje szczególnie chronione zależnie od celu i kontekstu. Wymagane są: analiza podstawy prawnej, minimalizacja zakresu danych, kontrola retencji, uprawnień, anonimizacja lub pseudonimizacja oraz rozdzielenie badań eksperymentalnych od automatycznego profilowania osób.

Rozporządzenie UE 2024/1689 (AI Act) w art. 5 ust. 1 lit. f zakazuje wnioskowania o emocjach osób w obszarze miejsca pracy i instytucji edukacyjnych, z wyjątkiem zastosowań ze względów medycznych lub bezpieczeństwa. Zakazy te zaczęły obowiązywać 2 lutego 2025 r.; pozostałe obowiązki wdrażano etapami. Kwalifikacja prawna konkretnego systemu zależy od definicji systemu rozpoznawania emocji, danych biometrycznych, rzeczywistego celu i środowiska użycia. Nie należy interpretować wyjątków jako blankietowego zezwolenia na profilowanie psychologiczne [11,12].

Badania dotyczące bólu, stresu czy obciążenia poznawczego wymagają szczególnej ostrożności w oznaczaniu wyjścia. Na przykład pomiar ruchów twarzy, zmęczenia lub fizjologicznego pobudzenia nie musi być tożsamy z prawnie

i metodologicznie rozumianym wnioskowaniem o emocjach. Każdą aplikację należy kwalifikować na podstawie jej rzeczywistego działania i obowiązujących przepisów, a nie samej nazwy modułu.

13. Ryzyka, luki i priorytety badawcze

Ryzyko	Skutek	Zabezpieczenie
Nadmierna interpretacja twarzy	Błędne wnioskowanie o emocji	Opis AU, kontekst, abstencja, walidacja ekspercka.
Leakage między osobami	Zawyżona skuteczność	LOSO/group split i audyt zbiorów.
Artefakty EEG/PPG/EDA	Fałszywe korelacje	Kontrola jakości, odrzucanie, cechy ruchu.
Niespójność czasowa	Pozorna utrata/zmiana dynamiki	Timestampy, ocena dryfu, analiza opóźnienia.
Braki modalności	Niestabilna fuzja	Maskowanie, fallback, kalibracja confidence.
Zmiana domeny	Spadek jakości in-the-wild	Walidacja zewnętrzna i podgrupy.
Kwantyzacja/pruning	Nierówna degradacja grup	Ablacje po eksporcie i na sprzęcie.
Nieuprawnione zastosowanie	Ryzyko prawne i etyczne	Analiza przeznaczenia i zgodności z AI Act/RODO.

14. Wnioski

Wartość badań nad analizą afektywną nie wynika z samej liczby klas ani z najwyższej wartości accuracy, lecz z możliwości odtworzenia wyników, pomiaru jakości anotacji, odporności na zmianę osoby i otoczenia oraz rozróżnienia zachowania ekspresyjnego od subiektywnego afektu. Architektury oparte na uwadze i fuzji multimodalnej stanowią uzasadnione hipotezy rozwojowe, ale ich przewagę trzeba wykazać w kontrolowanych ablacyjnych porównaniach. Dla pracy na Edge AI niezbędny jest wspólny pomiar jakości inferencji, kosztu pamięci, latencji end-to-end, energii i częstości utraty obserwacji. Decyzje o dopuszczalnym zastosowaniu muszą uwzględniać nie tylko osiągi klasyfikatora, ale także skutki społeczne, prywatność i ograniczenia regulacyjne.

15. Szczegółowa specyfikacja odtwarzalnego badania multimodalnego

15.1. Definicja zadania i ontologia stanów

Przed rozpoczęciem rejestracji należy opisać każdą zmienną wynikową poprzez konstrukcję teoretyczną, źródło etykiety, skalę oraz okres, do którego etykieta się odnosi. Dla ekspresji twarzy mogą to być jednostki FACS; dla sygnałów samoopisowych, walencja i pobudzenie wyrażane na wcześniej ustalonej skali; dla sygnałów autonomicznych, parametry fizjologiczne bez jednoznacznego przypisania do kategorii emocjonalnych. Zmiana rodzaju etykiety między eksperymentami bez jawnego rozróżnienia powoduje, że modele nie rozwiązują tego samego zadania.

Zaleca się stworzenie słownika metadanych obejmującego: identyfikator badania, pseudonim uczestnika, identyfikator sesji, bodźca, stan zgody, modalność, częstotliwość próbkowania, rodzaj aparatury, referencję EEG, typ filtru, jakość sygnału, procedurę anotacji i wersję etykiet. Każda transformacja powinna zachować identyfikowalny związek z nieprzetworzonym sygnałem w bezpiecznym archiwum badawczym; publiczne wyniki i modele nie muszą zawierać informacji umożliwiających identyfikację osoby.

15.2. Synchronizacja modalności i fragmentacja danych

Obraz, mowa i sygnały fizjologiczne pracują na różnych częstotliwościach. Kamera może generować klatki ze zmiennym odstępem, audio może być próbkowane w kilkunastu lub kilkudziesięciu kilohercach, a EEG i PPG wymagają niezależnych zegarów urządzeń. Przed segmentacją należy wyznaczyć wspólną oś czasu i określić opóźnienie każdego toru. Dopuszczalne są znaczniki sprzętowe, sygnały synchronizacyjne lub oszacowanie przesunięć zegara, lecz metoda i wynik muszą być raportowane.

Okna czasowe należy definiować względem początku i końca bodźca, a nie jedynie indeksów próbek. Przy nakładających się oknach zachodzi silna korelacja sąsiednich obserwacji; wszystkie okna pochodzące od tej samej sesji pozostają w jednym foldzie walidacyjnym. W protokole należy zdefiniować postępowanie z brakującymi klatkami, przerwami EEG, błędnymi

interwałami sercowymi i zerwaną synchronizacją. Sztuczne interpolowanie długich luk bez oznaczenia maski braków zwiększa ryzyko pozornych wzorców dynamiki.

15.3. Jakość sygnału i artefakty

Modalność	Zalecana kontrola jakości	Postępowanie
Wideo	Liczba wykrytych twarzy, orientacja, rozmycie, zasłonięcie	Odrzucenie lub oznaczenie okna niskiej jakości.
Audio	SNR, przesterowanie, pogłos, cisza, liczba mówców	Maskowanie i jawna informacja o niedostępności toru.
EEG	Kanały odstające, mrugnięcia, EMG, drifty i zanik kontaktu	Oznaczenie artefaktów, kontrola filtracji i referencji.
ECG / PPG	Artefakty ruchowe, brak impulsów, błędne interwały	Odrębny wskaźnik jakości przed analizą HRV.
EDA	Artefakty ruchowe, temperatura, kontakt elektrody	Podział komponentu tonic/fazowego i oznaczenie zmian bazy.
Tekst	Błędy ASR, niezrozumiałość, rozbieżność językowa	Walidacja transkrypcji i analiza wpływu ASR.

15.4. Strategia uczenia i walidacji międzyosobniczej

Dla badań zmierzających do wnioskowania o nowych osobach rekomendowany jest schemat leave-one-subject-out (LOSO) lub grupowa walidacja krzyżowa. Hiperparametry, normalizację cech, selekcję kanałów i próg decyzji należy dobierać wyłącznie wewnątrz zbioru treningowego, stosując nested validation, jeżeli liczba dostępnych grup umożliwia taki podział. Użytkownicy ze zbioru testowego nie mogą uczestniczyć w personalizacji, chyba że jest ona jawnie zdefiniowaną interwencją badawczą ocenianą na osobnym zbiorze hold-out.

Wartości niepewności należy szacować na poziomie niezależnych osób albo sesji, a nie pojedynczych okien, ponieważ liczba okien sztucznie podnosi efektywną liczebność. Dla porównania dwóch modeli wykorzystujących tych samych uczestników powinien być raportowany rozkład sparowanych różnic. Wyniki rozstrzygające powinny obejmować co najmniej medianę, przedział ufności, liczbę uczestników w teście i bezwzględną oraz względną zmianę jakości.

15.5. Kalibracja i interpretowalność predykcji

Rozpoznawanie emocji często jest oceniane po twardej decyzji argmax, co usuwa informacje o niepewności.
W zastosowaniach HCI korzystne jest raportowanie rozkładów wynikowych, jakości kalibracji (np. ECE, reliability diagram), odsetka predykcji z abstencją oraz skuteczności w funkcji progu ufności. W modelach wielomodalnych nie wystarcza wizualizacja wag attention jako rzekomego wyjaśnienia przyczynowego; należy przeprowadzać testy usuwania modalności i analizy kontrfaktyczne na danych wejściowych.

Przypadki rozbieżności, na przykład uśmiech i wysoka reakcja autonomiczna bez jednoznacznej samooceny, powinny pozostawać rozbieżnościami pomiarowymi, nie zaś być automatycznie redukowane do arbitralnej kategorii emocji. Dla zastosowań medycznych lub psychologicznych konieczna jest odrębna analiza trafności kryterialnej i konsekwencji błędu.

16. Eksperymenty ablacyjne i testy wdrożenia Edge AI

Eksperyment	Zmienna kontrolowana	Wynik do raportowania
Unimodal vs fuzja	Te same osoby, okna i etykiety	Δmacro-F1 / ΔCCC z CI.
Brak modalności	Celowo wyłączony jeden tor	Degradacja jakości i odsetek abstencji.
Szum / artefakty	SNR, poziom ruchu, zasłonięcie	Krzywa odporności i failure rate.
Kwantyzacja	FP32 → FP16 → INT8	Jakość, P95 latencji, RAM i energia.
Zmiana domeny	Nowa kamera, mikrofon, osoba lub sesja	Wynik OOD i błąd kalibracji.
Długi przebieg	Czas pracy i temperatura	Throttling, jitter, restart, ubytek danych.
Prywatność	Dane na urządzeniu / transmisja	Zakres retencji i skuteczność segmentacji.

Przewidywane usprawnienia po kwantyzacji lub przeniesieniu modelu na NPU nie są dowodami wydajności. Należy badać finalny binarny artefakt, z rzeczywistym kodem detekcji i filtracji, uruchomiony przy typowym obciążeniu urządzenia. Poprawa czasu inferencji przy jednoczesnym wzroście utraty ramek może pogorszyć całkowitą jakość systemu.

17. Minimalny standard publikowania wyników

Każdy wynik liczbowy powinien być powiązany z konkretną wersją zbioru, liczbą osób, modalnością, definicją etykiety, podziałem trening-test, preprocessingiem, architekturą, hiperparametrami, liczbą powtórzeń i miarą niepewności. Dla modeli raportowanych jedynie pojedynczą wartością accuracy bez tych pól należy zamieszczać informację „wynik wymagający weryfikacji protokołu”. Wniosek o przewadze technologii Edge AI wymaga równoczesnego wykazania jakości modelu, czasu i kosztu energetycznego, a wniosek o praktycznej przydatności w interakcji wymaga walidacji z udziałem rzeczywistych użytkowników.

Bibliografia

[1] Picard RW. Affective Computing. MIT Press; 1997.

[2] Ekman P, Friesen WV, Hager JC. Facial Action Coding System. Research Nexus; 2002.

[3] Russell JA. A circumplex model of affect. Journal of Personality and Social Psychology. 1980;39(6):1161–1178.

[4] Barrett LF, Adolphs R, Marsella S, Martinez AM, Pollak SD. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. Psychol Sci Public Interest. 2019;20(1):1–68. doi:10.1177/1529100619832930.

[5] Koelstra S et al. DEAP: A Database for Emotion Analysis Using Physiological Signals. IEEE Trans Affective Computing. 2012;3(1):18–31.

[6] Miranda-Correa JA et al. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. IEEE Trans Affective Computing. 2021;12(2):479–493.

[7] Schmidt P et al. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. ICMI; 2018.

[8] Mollahosseini A, Hasani B, Mahoor MH. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. IEEE Trans Affective Computing. 2019.

[9] Poria S et al. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. ACL; 2019.

[10] Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). PLOS ONE. 2018;13:e0196391.

[11] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2024/1689 ustanawiające zharmonizowane przepisy dotyczące sztucznej inteligencji; tekst skonsolidowany: https://eur-lex.europa.eu/eli/reg/2024/1689

[12] Komisja Europejska. Navigating the AI Act. https://digital-strategy.ec.europa.eu/en/faqs/navigating-ai-act

[13] NVIDIA DeepStream, Intel OpenVINO, Rockchip RKNN Toolkit: dokumentacja techniczna środowisk wdrożeniowych (specyfikacje wersji wymagane przy odtwarzaniu eksperymentu).